Machine Learning 2 - Homework Assignment 2

Alexandra Lindt

October 2, 2019

Problem 1

1.

The mutual information of two random variables X and Y with corresponding probability distributions $p(X)$ and $p(Y)$ is defined in Bishop equation 1.120 as the Kullback-Leibler divergence between the product of the variables distributions, i.e. $p(X) \cdot p(Y)$, and their joint distribution $p(X, Y)$.

$$\begin{aligned} \mathrm{I}[x, y] &\equiv \mathrm{KL}(p(x, y) \| p(x)p(y)) \\ &= -\iint p(x, y) \ln \left(\frac{p(x)p(y)}{p(x, y)} \right) \mathrm{d}x\mathrm{d}y \end{aligned}$$

From this we can deduce that $\mathrm{I}[x, y] \geq 0$, where $\mathrm{I}[x, y] = 0$ if and only if X and Y are independent with $p(x) \cdot p(Y) = p(X, Y)$. In case that X and Y are dependent, the mutual information is a measure of how similar the distributions $p(x) \cdot p(Y)$ is to $p(X, Y)$ and therefore of how strong their dependency is (the higher $\mathrm{I}[x, y]$, the 'more dependent' X and Y). Another interpretation of the mutual information p(follows from its relation to the conditional entropy.

$$\mathrm{I}[x, y] = \mathrm{H}[x] - \mathrm{H}[x|y] = \mathrm{H}[y] - \mathrm{H}[y|x]$$

With this we can interpret the mutual information $\mathrm{I}[x, y]$ as the reduction of our uncertainty about X if we had observed Y and vice versa.

Adding a new random variable Z with distribution $p(Z)$, we can define the conditional mutual information $\mathrm{I}[x, y|z]$ as

$$\begin{aligned} \mathrm{I}[x, y|z] &= \mathrm{H}[x|z] - \mathrm{H}[x|y, z] \\ &= \int p(z) \mathrm{KL}(p(x, y|z) \| p(x|z)p(y|z)) \mathrm{d}z \\ &= -\iiint p(z)p(x, y|z) \ln \left(\frac{p(x|z)p(y|z)}{p(x, y|z)} \right) \mathrm{d}x\mathrm{d}y\mathrm{d}z \end{aligned}$$

As we can see, the conditional mutual information is the expected value of the mutual information between two random variables X and Y when the value of another random variable Z is observed. If X and Y are both independent of Z, then $\mathrm{I}[x, y] = \mathrm{I}[x, y|z]$..

2.

As suggested, we first calculate the distributions.

$$p(x,y) = \sum_z p(x,y,z) \qquad p(x) = \sum_y p(x,y) \qquad p(y) = \sum_x p(x,y)$$

$$\Longrightarrow$$

x	y	$p(x,y)$	$p(x)$	$p(y)$
0	0	0.336	0.6	0.592
0	1	0.264	0.6	0.408
1	0	0.256	0.4	0.592
1	1	0.144	0.4	0.408

Since the variable's distributions are discrete, computing the mutual information reduces to

$$\begin{aligned} \mathrm{I}[x,y] &= -\sum_x \sum_y p(x,y) \ln \left(\frac{p(x)p(y)}{p(x,y)} \right) \\ &= \sum_x \sum_y p(x,y) \ln \left(\frac{p(x,y)}{p(x)p(y)} \right) \end{aligned}$$

Inserting the results from the above table into this formula, we receive

$$\mathrm{I}[x,y] \approx 0.0032$$

Since $\mathrm{I}[x,y] > 0$, X and Y are indeed dependent.

3.

Again, the distributions are calculated first.

$$p(z) = \sum_x \sum_y p(x,y,z) \quad p(x,y|z) = \frac{p(x,y,z)}{p(z)} \quad p(x|z) = \sum_y p(x,y|z) \quad p(y|z) = \sum_x p(x,y|z)$$

$$\Longrightarrow$$

x	y	z	$p(z)$	$p(x,y\|z)$	$p(x\|z)$	$p(y\|z)$
0	0	0	0.48	0.4	0.5	0.8
0	0	1	0.52	0.277	0.692	0.4
0	1	0	0.48	0.1	0.5	0.2
0	1	1	0.52	0.415	0.692	0.6
1	0	0	0.48	0.4	0.5	0.8
1	0	1	0.52	0.123	0.307	0.4
1	1	0	0.48	0.1	0.5	0.2
1	1	1	0.52	0.185	0.307	0.6

Using this result as well as the result from the previous sub-problem, we can calculate

$$\mathrm{I}[x,y|z] = \sum_z \sum_x \sum_y p(z)p(x,y|z) \ln \left(\frac{p(x,y|z)}{p(x|z)p(y|z)} \right) = 0$$

This means that X and Y are independent given that Z is observed.

4.

$$\begin{aligned} p(x,y,z) &= p(y,z|x)p(x) \\ &= p(y|z,x)p(z|x)p(x) \\ \star &= p(y|z)p(z|x)p(x) \end{aligned}$$

$\star$: Where we wrote $p(y|z,x) = p(y|z)$, because we know from the previous sub-problem that X and Y are independent when Z is given. The directed graph corresponding to this is depicted in figure 1.

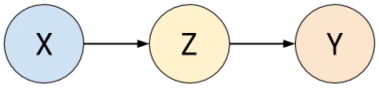

Figure 1: Corresponding directed graph.

Problem 2

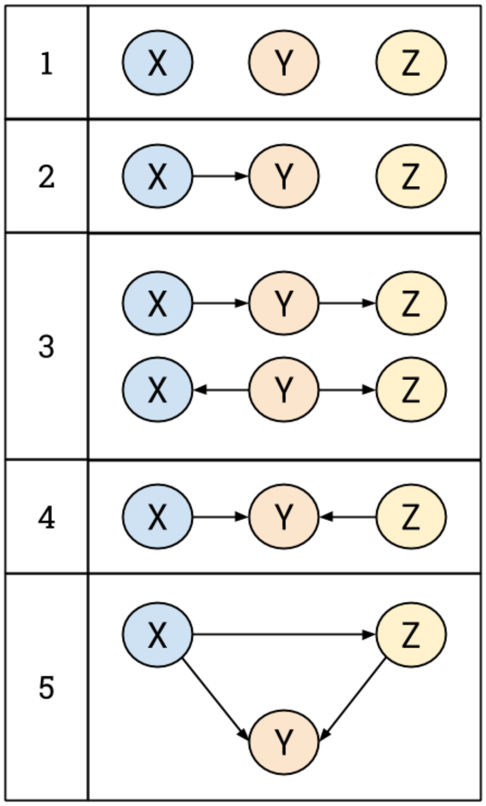

Figure 2: Five clusters of all bayesian networks with three vertices X, Y and Z.

Considering all bayesian networks consisting of three vertices, we find that they can be grouped into the five different clusters depicted in figure 2. The graphs of each cluster encode the same set of independence relations between the variables. Note that any graphs that could be transformed into one of the depicted graphs by swapping the variable's positions in the graph structure are not included. The independence relations for each cluster are given in the following table. Note that the clusters are numbered as in figure 2.

	Cluster Number	**Independence Relations**
	1	$\mathrm{X} \perp\!\!\!\perp Y \quad X \perp\!\!\!\perp Y \mid Z$ $\mathrm{X} \perp\!\!\!\perp Z \quad X \perp\!\!\!\perp Z \mid Y$ $\mathrm{Y} \perp\!\!\!\perp Z \quad Y \perp\!\!\!\perp Z \mid X$
	2	$\mathrm{X} \not\!\perp\!\!\!\perp Y \quad X \not\!\perp\!\!\!\perp Y \mid Z$ $\mathrm{X} \perp\!\!\!\perp Z \quad X \perp\!\!\!\perp Z \mid Y$ $\mathrm{Y} \perp\!\!\!\perp Z \quad Y \perp\!\!\!\perp Z \mid X$
$\Longrightarrow$	**3**	$\mathrm{X} \not\!\perp\!\!\!\perp Y \quad X \not\!\perp\!\!\!\perp Y \mid Z$ $\mathrm{X} \not\!\perp\!\!\!\perp Z \quad X \perp\!\!\!\perp Z \mid Y$ $\mathrm{Y} \not\!\perp\!\!\!\perp Z \quad Y \not\!\perp\!\!\!\perp Z \mid X$
	4	$\mathrm{X} \not\!\perp\!\!\!\perp Y \quad X \not\!\perp\!\!\!\perp Y \mid Z$ $\mathrm{X} \not\!\perp\!\!\!\perp Z \quad X \not\!\perp\!\!\!\perp Z \mid Y$ $\mathrm{Y} \perp\!\!\!\perp Z \quad Y \not\!\perp\!\!\!\perp Z \mid X$
	5	$\mathrm{X} \not\!\perp\!\!\!\perp Y \quad X \not\!\perp\!\!\!\perp Y \mid Z$ $\mathrm{X} \not\!\perp\!\!\!\perp Z \quad X \not\!\perp\!\!\!\perp Z \mid Y$ $\mathrm{Y} \not\!\perp\!\!\!\perp Z \quad Y \not\!\perp\!\!\!\perp Z \mid X$

Problem 3

1.

$$
\begin{aligned}
KL(p\|q) &= -\int p(x)\log\left(\frac{q(x)}{p(x)}\right)dx \\
&= \int \big(\log(p(x)) - \log(q(x))\big)p(x)dx \\
&= \int \left(\log\left(\frac{|L|^{1/2}}{|\Sigma|^{1/2}}\right) - \frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) + \frac{1}{2}(x-m)^T L^{-1}(x-m)\right)p(x)dx \\
&= \int \log\left(\frac{|L|^{1/2}}{|\Sigma|^{1/2}}\right)p(x)dx - \frac{1}{2}\int (x-\mu)^T\Sigma^{-1}(x-\mu)p(x)dx + \frac{1}{2}\int (x-m)^T L^{-1}(x-m)p(x)dx \\
\star &= \frac{1}{2}\left(\log\left(\frac{|L|}{|\Sigma|}\right) - (\mu-\mu)^T\Sigma^{-1}(\mu-\mu) - \mathrm{Tr}(\Sigma^{-1}\Sigma) + (\mu-m)^T L^{-1}(\mu-m) + \mathrm{Tr}(\text{Ł}^{-1}\Sigma)\right) \\
&= \frac{1}{2}\left(\log\left(\frac{|L|}{|\Sigma|}\right) - \mathrm{Tr}(I_D) + (\mu-m)^T L^{-1}(\mu-m) + \mathrm{Tr}(\text{Ł}^{-1}\Sigma)\right) \\
&= \frac{1}{2}\left(\log\left(\frac{|L|}{|\Sigma|}\right) - D + (\mu-m)^T L^{-1}(\mu-m) + \mathrm{Tr}(\text{Ł}^{-1}\Sigma)\right)
\end{aligned}
$$

$\star$: Here we make use of formula 380 on page 43 of the matrix cookbook

$$\begin{aligned} E[(x-m')^T A(x-m')] &= \int p(x)(x-m')^T A(x-m')dx \\ &= (m-m')^T A(m-m') + \operatorname{Tr}(A\Sigma) \end{aligned}$$

where m denotes the mean and Σ the variance of the probability distribution $p(x)$.

2.

$$\begin{aligned} H(p) &= -\int p(x)\log p(x)dx \\ &= -\int p(x)\Big(\log(\frac{1}{(2\pi)^{\frac{D}{2}}}) + \log(\frac{1}{|\Sigma|^{\frac{1}{2}}}) - \frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\Big)dx \\ &= \int p(x)(\frac{D}{2}\log(2\pi) + \frac{1}{2}\log(|\Sigma|)\big)dx + \frac{1}{2}\int p(x)(x-\mu)^T\Sigma^{-1}(x-\mu)dx \\ \star &= \frac{D}{2}\log(2\pi) + \frac{1}{2}\log(|\Sigma|) + \frac{1}{2}(\mu-\mu)^T\Sigma^{-1}(\mu-\mu) + \frac{1}{2}Tr(\Sigma^{-1}\Sigma) \\ &= \frac{1}{2}\Big(D\log(2\pi) + \log(|\Sigma|) + D\Big) \end{aligned}$$

$\star$: Here we do again make use of the matrix cookbook formula 380 given in the $\star$-explanation of the previous sub-problem.